

---

# Root-cause analysis for time-series anomalies via spatiotemporal causal graphical modeling

---

**Chao Liu**

Department of Mechanical Engineering  
Iowa State University  
Ames, IA 50010  
cliu5@iastate.edu

**Kin Gwn Lore**

Department of Mechanical Engineering  
Iowa State University  
Ames, IA 50010  
kglore@iastate.edu

**Soumik Sarkar**

Department of Mechanical Engineering  
Iowa State University  
Ames, IA 50010  
soumiks@iastate.edu

## Abstract

Modern distributed cyber-physical systems encounter a large variety of anomalies and in many cases, they are vulnerable to catastrophic fault propagation scenarios due to strong connectivity among the sub-systems. In this regard, root-cause analysis becomes highly intractable due to complex fault propagation mechanisms in combination with diverse operating modes. This paper presents a new data-driven framework for root-cause analysis for addressing such issues. The framework is based on a spatiotemporal feature extraction scheme for multivariate time series built on the concept of symbolic dynamics for discovering and representing causal interactions among subsystems of a complex system. We propose sequential state switching ( $S^3$ ) and artificial anomaly association ( $A^3$ ) methods to implement root-cause analysis in an unsupervised and semi-supervised manner respectively. Synthetic data from cases with failed pattern(s) and anomalous node are simulated to validate the proposed approaches, then compared with the performance of vector autoregressive (VAR) model-based root-cause analysis. The results show that: (1)  $S^3$  and  $A^3$  approaches can obtain high accuracy in root-cause analysis and successfully handle multiple nominal operation modes, and (2) the proposed tool-chain is shown to be scalable while maintaining high accuracy.

## 1 Introduction

With the advent ubiquitous sensing, advanced computation and strong connectivity, modern distributed cyber-physical systems (CPSs) such as power plants, integrated buildings, transportation networks and power-grids have shown tremendous potential of increased efficiency, robustness and resilience. From the perspective of performance monitoring, anomaly detection and root-cause analysis of such systems, technical challenges arise from a large number of subsystems that are highly interactive and operate in diverse modes.

For the purpose of root-cause analysis for time-series anomalies in complex systems, Granger causality is applied to model the system-wide behavior and capture the variation that can be used to implement root-cause analysis. With multivariate time series data, studies show that the causality from and to the fault variable presents differences and can be used to reason the root-cause [1–4]. For anomaly detection in time series, Qiu et. al. [5] derived neighborhood similarity and coefficient similarity from Granger-Lasso algorithm, to compute anomaly score and ascertain threshold for

anomaly detection. A causality analysis index based on dynamic time warping is proposed by Li et. al. [6] to determine the causal direction between pairs of faulty variables in order to overcome the shortcoming of Granger causality in nonstationary time series. Fault related variables can be clustered and root-cause analysis is implemented in each cluster. A dynamic uncertain causality graph is proposed for probabilistic reasoning and applied to fault diagnosis of generator system of nuclear power plant [7]. Also, Bayesian network has been applied in cyber-physical systems to implement anomaly detection and root-cause analysis [8]. The proposed approaches provide efficient tools in discovering causality in complex systems, while an approach in inferencing (interpreting the variation in causality into decisions on failed patterns of fault variable/node) is less investigated.

In this context, we present a semi-supervised tool for root-cause analysis in complex systems based on a data driven framework proposed for system-wide time-series anomaly detection in distributed complex system [9], and using a spatiotemporal feature extraction scheme built on the concept of symbolic dynamics for discovering and representing causal interactions between the subsystems. The proposed tool aims to (i) capture multiple operational modes as nominal in complex CPSs, (ii) only use nominal data and artificially generated fault data to train the model without requiring true labeled anomaly data, and (iii) implement root-cause analysis in a semi-supervised way in a diversity of faults (e.g., one failed pattern, multiple failed patterns, one fault node, and multiple fault nodes). We present two approaches for root-cause analysis, namely the *sequential state switching* ( $S^3$ , based on free energy concept of a Restricted Boltzmann Machine, RBM [10]) and *artificial anomaly association* ( $A^3$ , a multi-label classification framework using deep neural networks, DNN). Synthetic data from cases with failed pattern(s) and faulty node are simulated to validate the proposed approaches, then compared with the vector autoregressive (VAR) model in terms of root-cause analysis performance.

## 2 Background and preliminaries

### 2.1 Spatiotemporal pattern network (STPN)

STPN modeling involves partitioning/discretization, followed by learning markov machines. While details can be found in [11, 9], we are providing brief description for completeness.

Consider a multivariate time series,  $X = \{X^{\mathbb{A}}(t), t \in \mathbb{N}, \mathbb{A} = 1, 2, \dots, n\}$ , where  $n$  is the number of variables or dimension of the time series, corresponding to the number of nodes in graphical modeling. Let  $\mathbb{X}$  denote a set of partitioning/discretization functions [12],  $\mathbb{X} : X(t) \rightarrow S$ , that transform a general dynamic system (time series  $X(t)$ ) into a symbol sequence  $S$  with an alphabet set  $\Sigma$ . Various partitioning approaches have been proposed in the literature, such as uniform partitioning (UP), maximum entropy partitioning (MEP, used for the present study), maximally bijective discretization (MBD) [13] and statistically similar discretization (SSD) [14]. Subsequently, a probabilistic finite state automaton (PFSA) is defined to describe states (representing various parts of the data space) and probabilistic transitions among them (can be learnt from data) via  $D$ -Markov machine and  $xD$ -Markov machine. With this setup, an STPN is defined as:

**Definition.** A PFSA based STPN is a 4-tuple  $W_D \equiv (Q^a, \Sigma^b, \Pi^{ab}, \Lambda^{ab})$ : ( $a, b$  denote nodes of the STPN)

1.  $Q^a = \{q_1, q_2, \dots, q_{|Q^a|}\}$  is the state set corresponding to symbol sequences  $S^a$ .
2.  $\Sigma^b = \{\sigma_0, \dots, \sigma_{|\Sigma^b|-1}\}$  is the alphabet set of symbol sequence  $S^b$ .
3.  $\Pi^{ab}$  is the symbol generation matrix of size  $|Q^a| \times |\Sigma^b|$ , the  $ij^{th}$  element of  $\Pi^{ab}$  denotes the probability of finding the symbol  $\sigma_j$  in the symbol string  $s^b$  while making a transition from the state  $q_i$  in the symbol sequence  $S^a$ ; while self-symbol generation matrices are called atomic patterns (APs) i.e., when  $a = b$ , cross-symbol generation matrices are called relational patterns (RPs) i.e., when  $a \neq b$ .
4.  $\Lambda^{ab}$  denotes a metric that can represent the importance of the learnt pattern (or degree of causality) for  $a \rightarrow b$  which is a function of  $\Pi^{ab}$ .

### 2.2 Unsupervised anomaly detection with spatiotemporal causal graphical modeling

A data-driven framework for system-wide anomaly detection is proposed in [9], noted as the **STPN+RBM** model. The steps of learning the STPN+RBM model are:

1. Learn APs and RPs (individual node behaviors and pair-wise interaction behaviors) from the multivariate training symbol sequences.

2. Consider short symbol sub-sequences from the training sequences and evaluate  $\Lambda^{ij} \forall i, j$  for each short sub-sequence.
3. For one sub-sequence, based on a user-defined threshold on  $\Lambda^{ij}$ , assign state 0 or 1 for each AP and RP; thus every sub-sequence leads to a binary vector of length  $L$ , where  $L = \#AP + \#RP$ .
4. An RBM is used for modeling system-wide behavior with nodes in the visible layer corresponding to APs and RPs.
5. The RBM is trained using binary vectors generated from nominal training sub-sequences.
6. Online anomaly detection is implemented by computing the probability of occurrence of a test STPN pattern vector via trained RBM (as shown in Fig. 1).

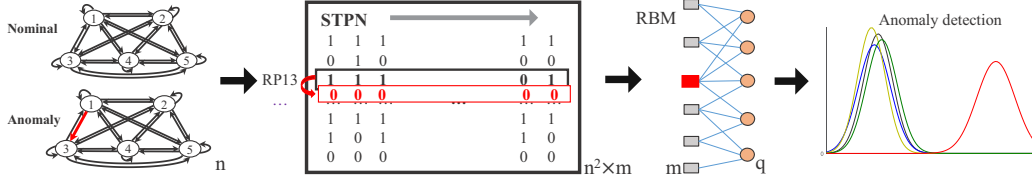


Figure 1: Anomaly detection process. Spatiotemporal features are extracted from both nominal and anomalous data, with online STPN applied. Multiple sub-sequences of APs and RPs form input vectors to the RBM. Here, the RBM is only trained with nominal data, and the anomalous data is used as input to compute free energy. Anomaly is detected by identifying its high energy state, which can be captured using the Kullback-Leibler Distance (KLD) metric [9].

### 3 Methods

In the STPN+RBM framework described above, anomaly manifests itself as a low probability (high energy) event. Therefore, the idea for root-cause analysis is to find potential pattern(s) that, if changed, can transition the system from a high to a low energy state. The probabilities of AP and RP's existence are discovered by the STPN, and an anomaly will influence the causality of specific patterns (e.g., in STPN, the probability of the pattern might be switched/flipped from 0 to 1). Hence, by switching/flipping a pattern, its contribution on the energy states of the system can be identified and a large contribution may indicate the root-cause of an anomaly. Using this principle, two approaches—the sequential state switching ( $S^3$ ) and artificial anomaly association ( $A^3$ )—are proposed.

#### 3.1 Sequential state switching ( $S^3$ )

For an  $n$ -node graphical model, all the APs and RPs together form a binary vector  $v$  of length  $L = n^2$  ( $L = \#AP + \#RP$ , where  $\#AP = n$ ,  $\#RP = n \times (n - 1)$ ). One such binary vector is treated as one training example for the system-wide RBM (with  $n^2$  number of visible units) and many such examples are generated from different short sub-sequences extracted from the overall training sequence. Then, the RBM is trained by maximizing the maximum likelihood of the data.

During training, weights and biases are obtained such that the training data has low energy. During inference then, an anomalous pattern should manifest itself as a low probability (high energy) configuration. The energy function for an RBM is defined as:

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{h}^T \mathbf{W} \mathbf{v} - \mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h}$$

where  $\mathbf{W}$  are the weights of the hidden units,  $\mathbf{b}$  and  $\mathbf{c}$  are the biases of the visible units and hidden units respectively.

With the weights and biases of RBM, free energy can be computed. Free energy is defined as the energy that a single visible layer pattern would need to have in order to have the same probability as all of the configurations that contain  $\mathbf{v}$  [15], which has the following expression:

$$F(v) = -\sum_i v_i a_i - \sum_j \log(1 + e^{b_j + \sum_i v_i w_{ij}})$$

The free energy in nominal conditions is noted as  $\tilde{F}$ . In cases where there are multiple input vectors with more than one nominal modes, free energy in the nominal states can be averaged or used in conjunction with other metrics. In anomalous conditions, a failed pattern will shift the energy from

a lower state to a higher state. Assume that the patterns can be categorized into two sets,  $\mathbf{v}^{nom}$  and  $\mathbf{v}^{ano}$ . By flipping the set of anomalous patterns  $\mathbf{v}^{ano}$ , a new expression for free energy is obtained:

$$F^s(v) = - \sum_g v_g a_g - \sum_j \log(1 + e^{b_j + \sum_g v_g w_{gj}}) \\ - \sum_h v_h^* a_h - \sum_j \log(1 + e^{b_j + \sum_h v_h^* w_{hj}}), \{v_g\} \in v^{nom}, \{v_h^*\} \in v^{*,ano}$$

Here,  $v^*$  has the opposite state to  $v$  and represents that the probability of the pattern has been significantly changed. In this work, the probabilities of the patterns are binary (i.e. 0 or 1). Hence, we have that  $v^* = 1 - v$ . The sequential state switching is formulated by finding a set of patterns  $v^{ano}$  via  $\min(F^s(v^{ano}, v^{nom}) - F)$ . Algorithm 1 is presented to perform root-cause analysis based on the STPN+RBM framework.

---

**Algorithm 1** Root-cause analysis with sequential state switching ( $S^3$ ) method

---

```

1: procedure STPN+RBM MODELING ▷ Algorithm 1 in [9]
2:   Online process of computing likelihoods/probabilities of APs & RPs
3:   Training RBM to achieve low energy state, using binary vectors from nominal subsequences
4: end procedure
5: procedure ANOMALY DETECTION ▷ Algorithm 2 in [9]
6:   Online anomaly detection via the probability of the current state via trained RBM
7: end procedure
8: procedure ROOT-CAUSE ANALYSIS
9:   if Anomaly = 1 then
10:     $F^c \leftarrow F^s(\mathbf{v})$  ▷  $F^c$  is the current free energy with input vector  $\mathbf{v} = \mathbf{v}^{nom} \cup \mathbf{v}^{ano}$ .
11:     $\{V_p\} \leftarrow \{v : F^s(\mathbf{v}) < F^c\}$ 
12:    while  $\{V_p\} \neq \emptyset \vee \{v : F^s(\mathbf{v}^{*,ano}, \mathbf{v}^{nom}) < F^c\} = \emptyset$  do
13:       $F^c \leftarrow \min(F^s(\mathbf{v}^{*,ano} \cup v_i^*, \mathbf{v}^{nom})), v_i \in \{V_p\}, v_i^* = 1 - v_i, \mathbf{v}^{*,ano} = 1 - \mathbf{v}^{ano}$ 
14:       $\{v^{ano}\} \leftarrow \{v^{ano}\} \cup \{v : F^s(\mathbf{v}^{*,ano}, \mathbf{v}^{nom}) = F^c\}$ 
15:       $\{v^{nom}\} \leftarrow \{v\} \setminus \{v^{ano}\}$ 
16:       $\{V_p\} \leftarrow \{V_p\} \setminus \{v : F^s(\mathbf{v}^{*,ano}, \mathbf{v}^{nom}) = F^c\}$ 
17:    end while
18:  end if
19:   $\{\Lambda^{ano}\} \leftarrow \{v^{ano}\}$ 
20:  return  $\{\Lambda^{ano}\}$ 
21: end procedure

```

---

It should be noted that free energy  $F$  is used in Algorithm 1, and it can be used along with other metrics such as KLD. Using KLD alongside with free energy is particularly useful when the distribution of free energy is obtained with multiple sub-sequences. KLD may be more robust as it takes multiple sub-sequences into account because a persistent anomaly across the subsequences will cause a significant impact on KLD.

### 3.2 Artificial anomaly association ( $A^3$ )

Using a deep neural network (DNN), we frame the root-cause analysis problem as a multi-class (binary-class in our case) classification problem [16]. The input is presented as an  $n^2$ -element vector with values of either 0 or 1 which denotes whether a specific pattern is activated. We desire to map the input vector to an output vector of the same length (termed as the *indicator label*), where the value of each element within the output vector indicates whether a specific pattern is anomalous. For nominal modes, the input vector may be comprised of different combinations of 0's and 1's, and the indicator labels will be a vector of all 1's (where the value 1 denotes no anomaly). However, if a particular element  $i$  within the input vector gets flipped, then the indicator label corresponding to the  $i$ -th position in the output vector will be flipped and switches from 1 (normal) to 0 (anomalous). In this way, we can identify that the  $i$ -th pattern is anomalous. With this setup, a classification sub-problem (i.e. is this pattern normal, or anomalous?) can be solved for each element in the output vector given the input data. One might argue that multi-class classification is unnecessary and adds to higher computational overhead. Although a single class denoting which pattern has failed may work for a single anomalous pattern case, it is not sufficient for simultaneous multiple pattern failures.

The crux of the idea is that anomalies are artificially injected into readily-available nominal data since anomalous data are not always available. A justification is that we wish the model to hierarchically extract important features from the nominal data to sufficiently discriminate between nominal and anomalous patterns. Furthermore, realistic physical systems may have multiple nominal nodes and it will most certainly be arduous to discover the root-cause using conventional methods when anomaly occurs. As deep learning methods are able to learn hierarchical representations of the input data, they are attractive candidates to solve the problem where multiple nominal modes are involved.

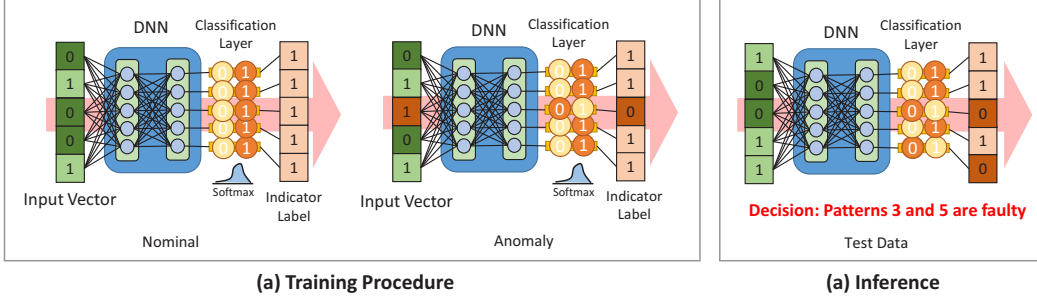


Figure 2: Framing the problem as an artificial anomaly association ( $A^3$ ) problem. (a) When training the model, the value of an element in both the input vector and its corresponding indicator label is randomly flipped to simulate anomaly. (b) When inferencing, a test input is fed into the DNN and a classification sub-problem is solved to obtain the indicator vector. Values of 0 in the output vector traces back to the exact patterns that are faulty.

**DNN Parameters:** Training data is generated from 6 modes including both nominal and artificial anomalous data. The training data is split into 222,300 training samples and 74,100 validation samples. Testing was performed on datasets generated from entirely different modes. The DNN is comprised of 3 layers of 500 hidden units each and trained with ReLU activations. A learning rate of 0.1 and a batch size of 10 is used in the gradient descent algorithm. The training procedure employs the early-stopping algorithm where training stops when validation error ceases to decrease.

## 4 Results and discussions

Synthetic datasets are generated with vector autoregressive (VAR) process to simulate anomaly in pattern(s)/node(s) for performance evaluation of  $S^3$  and  $A^3$  methods.

### 4.1 Anomaly in pattern(s)

**Dataset:** Anomaly in pattern(s) is defined as the change of one or more causal relationship, while defining anomaly, this translates to a changed/switched pattern in the context of STPN. A 5-node system is defined and shown in Fig. 3 including six different nominal modes. Cycles ( $1 \rightarrow 2 \rightarrow 5 \rightarrow 1$ ,  $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ ,  $2 \rightarrow 3 \rightarrow 2$ ) and loops ( $4 \rightarrow 4$ ) are included in the models. The graphical models are applied to simulate multiple nominal modes. Anomalies are simulated by breaking specific patterns in the graph; 30 cases are formed including 5 cases in one failed pattern, 10 cases in two failed patterns, 10 cases in three failed patterns, and 5 cases in four failed patterns. Multivariate time series data (denoted as dataset1) are generated using VAR process that follows the causality definition in the graphical models.

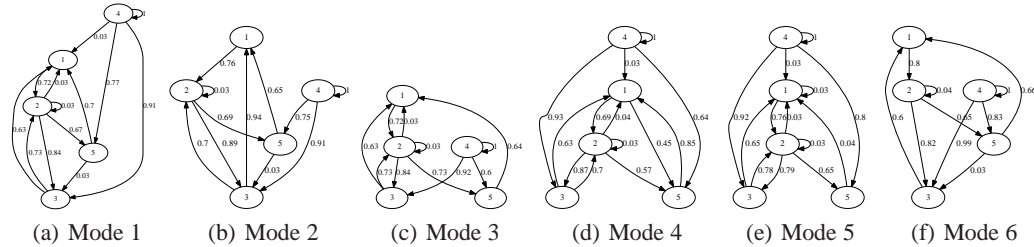


Figure 3: Graphical models defined to simulate pattern(s) anomaly. Six graphs are defined and treated as nominal operation modes in complex systems. Pattern failure is simulated by breaking specific patterns in the model (not shown).



**Performance Evaluation:** Root-cause identification performances of  $S^3$  and  $A^3$  methods are evaluated using dataset1. The accuracy metric used here is a pattern-by-pattern comparison (both anomalous and nominal) between the prediction and the ground truth labels. Formally, we define the accuracy metric  $\alpha_1$  as:

$$\alpha_1 = \frac{\sum_{j=1}^{n^2} \sum_{i=1}^m \chi_1(T_{ij} = P_{ij})}{mn^2}$$

where  $T_{ij}$  denotes the ground truth state (nominal/anomalous) of the  $j^{th}$  pattern of the  $i^{th}$  test sample. Similarly,  $P_{ij}$  is the corresponding predicted state using the root-cause analysis approach.  $\chi_1(\cdot)$  is the indicator function that returns the value 1 if the expression within the function is true. In the denominator,  $m$  is the number of test samples and  $n^2$  is the number of patterns. With the above metric, results of  $S^3$  method and  $A^3$  method are listed in Table 1. High *accuracy* is obtained for

Table 1: Root-cause analysis results in  $S^3$  method and  $A^3$  method with synthetic data.

Approach	Training samples	Testing samples	Accuracy $\alpha_1$ (%)
$S^3$	11,400	57,000	97.04
$A^3$	296,400	57,000	97.42

both  $S^3$  and  $A^3$  method. While training time is much less for  $S^3$ , the inference time in root-cause analysis for  $S^3$  is much more than that of  $A^3$ , as  $S^3$  depends on sequential searching. Note, the classification formulation in  $A^3$  aims to achieve the exact set of anomalous nodes. On the other hand, the  $S^3$  method is an approximate method that sequentially identifies anomalous patterns and hence, the stopping criteria would be critical. The observation that the performance of  $S^3$  is quite comparable to that of  $A^3$  suggests a reasonable choice of the stopping criteria.

It is evident from the above discussion that while the approximate  $S^3$  method may overestimate the set of anomalous patterns to some extent, it will be important to not miss any anomalous pattern. In this context, we define a more relaxed metric below to understand the efficacy of the  $S^3$  method.

$$\alpha_2 = \frac{\sum_{i=1}^m \chi_1(\{A_T\}_i \subseteq \{A_P\}_i)}{m}$$

where  $\{A_T\}_i$  denotes the ground truth set of anomalous patterns for test sample  $i$  and  $\{A_P\}_i$  denotes the corresponding predicted set. As expected, we obtain a higher accuracy value of 99.02% for this relaxed metric which shows that the approximate  $S^3$  approach is reliable in finding out all failed patterns.

## 4.2 Anomaly in node(s)

**Dataset:** Anomaly in node(s) occurs when one node or multiple nodes fail in the system. For a cyber-physical system, they may be caused by sensor fault or component degradation. The graphical model defined in Fig. 3 (a) is applied for generating the nominal data using the VAR process. Anomaly data are simulated by introducing time delay in a specific node. The time delay will break most of the causality to and from this node (except possibly the self loop, i.e., AP of the failed node). For validation and comparison, graphs are recovered with VAR process in cases of faulty nodes as shown in Fig. 4). The figure also shows that there are some variations in causality between normal nodes which suggests that causality discovery is more difficult with existence of cycles and loops. The generated dataset is denoted as dataset2. For scalability analysis, a 30-node system is further defined with VAR process, and the same method is applied to simulate the anomaly data, noted as dataset3.

**Performance Evaluation:** This work is aimed at discovering failed patterns instead of recovering underlying graph (which is very difficult in graphs with cycles and loops [17, 18]). The discovered anomalous patterns can then be used for diagnosing the fault node. For instance, a failed pattern  $N_i \rightarrow N_j$  discovered by root-cause analysis can be caused by the fault node  $i$  or  $j$ . However, if multiple failed patterns are related to the node  $i$ , then this node can be deemed anomalous. In this regard, it is important to learn the impact of one pattern on a detected anomaly compared to another. This can facilitate a ranking of the failed patterns and enable a robust isolation of an anomalous node. However, the exact  $A^3$  method does not provide such information and therefore, we investigate only results obtained with the sequential  $S^3$  method. Moreover, with anomaly injection process by introducing time delays at a node level prevents us from selecting an appropriate performance

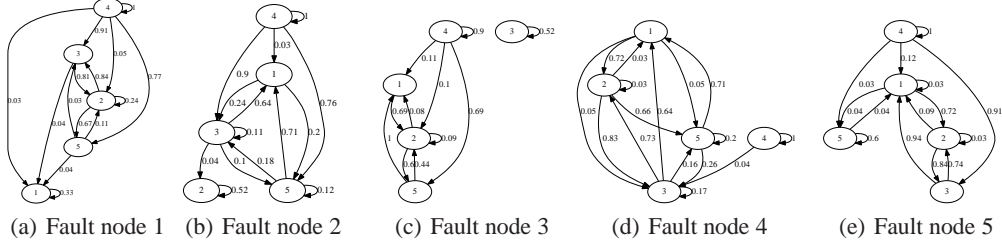


Figure 4: Anomalous conditions with fault node, causality is discovered by VAR model. Compared with the nominal mode in Fig. 3 (a), the fault node breaks most of the causality from and to this node. Causality is normalized by the maximal value among all of the patterns, and causality smaller than 0.03 is not shown.

metric for the  $A^3$  method. For comparison, we use VAR-based graph recovery method that is widely applied in economics and other sciences, and efficient in discovering Granger causality [19]. Note, the test dataset itself is synthetically generated using a VAR model with a specific time delay. Hence, the causality in such a multivariate time series is supposed to be well captured by VAR-based method. The details of the VAR-based root-cause analysis strategy is explained below.

Consider the time series  $Y(t) = y_{i,t}, i = (1, 2, \dots, n), t \in \mathbb{N}$

$$y_{i,t} = \sum_{k=1}^p \sum_{j=1}^n A_{i,j}^k y_{j,t-k} + \mu_t, \quad j = (1, 2, \dots, n)$$

where  $p$  is time lag order,  $A_{i,j}$  is the coefficient of the influence on  $i$ th time series caused by  $j$ th time series, and  $\mu_t$  is an error process with the assumption of white noise with zero mean, i.e.  $E(\mu_t) = 0$ , and that the covariance matrix  $E(\mu_t \mu_t') = \Sigma_\mu$  is time invariant.

With the given time series, a VAR model (i.e., the coefficients  $A_{i,j}$ ) can be learnt using standard algorithm [19]. The differences in coefficients between the nominal and anomalous models are subsequently used to find out the root causes. The pattern is deemed to have failed when  $\delta A_{i,j} > 0.4 \cdot \max\{\delta A_{i,j}\}$  where  $\delta A_{i,j} = |A_{i,j}^{ano} - A_{i,j}^{nom}|$ .

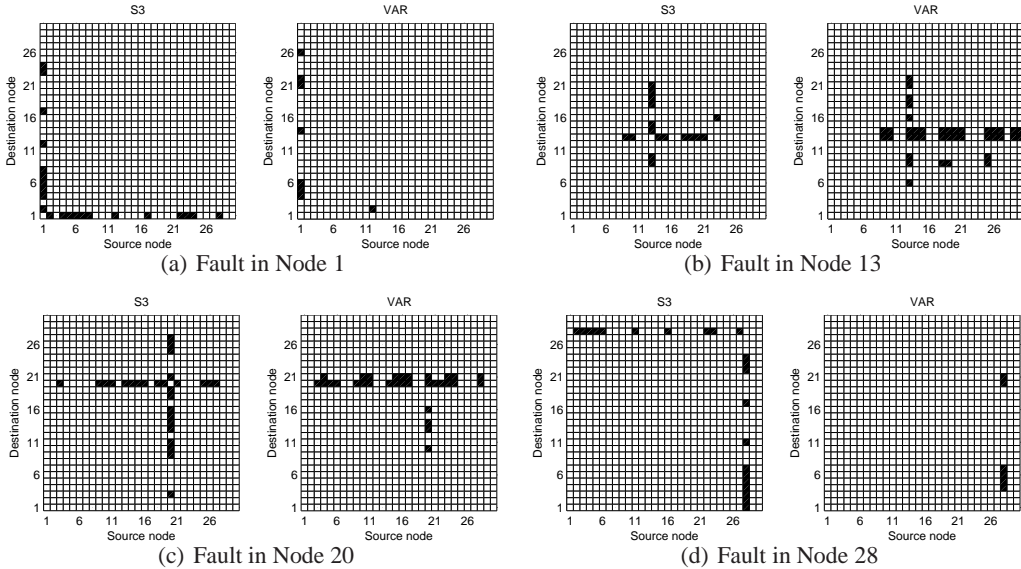


Figure 5: Comparisons between sequential state switching ( $S^3$ ) method and vector autoregressive (VAR) method using dataset3. The patterns represented with blocks are from the node in x-axis to the node in y-axis, and the discovered root causes are marked in black.

The results of  $S^3$  and VAR using dataset3 are shown in Fig. 5. In panel (a), all of the changed patterns discovered by  $S^3$  can be attributed to node 1 (shown by the black boxes in column and row 1). Therefore, node 1 is considered as faulty by the  $S^3$  method. On the other hand, VAR incorrectly

discovers a significant change in the pattern  $N_{12} \rightarrow N_2$  but not the patterns originating from  $N_1$ . We note this as an error. In general, although  $S^3$  and VAR can both discover the fault node, VAR produces more false alarms. Similar observations can be made in cases (b) and (c). In (d), 22 failed patterns are discovered by  $S^3$  which can all be attributed to a fault in node 28, but only six failed patterns are discovered by VAR. In this case, the discovery of more changed patterns can contribute to a stronger evidence that node 28 is faulty. As it is always beneficial to avoid false positive error, we conclude that  $S^3$  method is more preferred for case (d).

In real applications, when more anomalous patterns are discovered incorrectly, more effort will be needed to analyze the failed patterns closely and determine the root-cause node. This will lead to more financial expenditures and time investment in finding the failed node. With this motivation, an error metric  $\epsilon$  is defined by computing the ratio of incorrectly discovered anomalous patterns  $|\{\Lambda^\epsilon\}|$  to all discovered anomalous patterns  $|\{\Lambda^{ano}\}|$ , i.e.,  $\epsilon = \frac{|\{\Lambda^\epsilon\}|}{|\{\Lambda^{ano}\}|}$ . The results using dataset2 and dataset3 are listed in Table 2.

Table 2: Comparison of root-cause analysis results with  $S^3$  and VAR.

Approach	Dataset 2 (5 nodes)			Dataset 3 (30 nodes)		
	$ \{\Lambda^{ano}\} $	$ \{\Lambda^\epsilon\} $	$\epsilon$ (%)	$ \{\Lambda^{ano}\} $	$ \{\Lambda^\epsilon\} $	$\epsilon$ (%)
$S^3$	13	2	15.38	653	18	2.76
VAR	20	4	20.00	521	113	21.69

While it should be noted that both  $S^3$  and VAR can discover the fault node correctly in both the datasets, the error ratio for  $S^3$  method is much lower than that for VAR (i.e., lower false alarm). As the scale of the system increases, the number of discovered anomalous patterns by  $S^3$  becomes more than that of by VAR, while the error ratio becomes significantly less than that of VAR. Therefore,  $S^3$  method is both *scalable* as well as demonstrates better accuracy. Note that for comparisons between  $S^3$  and VAR, only one nominal mode is considered in Table 2 as VAR is not directly applicable in cases with multiple nominal modes.  $S^3$  method can handle multiple nominal modes and it has been validated in Section 4.1.

**Remark 4.1** *The root-cause analysis algorithms proposed in this work are fundamentally pattern-based as opposed to being node-based. Motivation for such methods come from real cyber-physical systems, where cyber-attacks may only compromise interactions among sub-systems (i.e., relational patterns (RPs)) without directly affecting sub-systems (i.e., the nodes). In contrast, majority of the existing methods mostly focus on node-based anomalies to the best of our knowledge. For example, in [5], root-cause analysis was performed using Granger Graphical Model with neighborhood similarity (GGM-N) and Granger Graphical Model with coefficient similarity (GGM-C). An initial investigation shows that for the single node failure cases in dataset3, GGM-N and GGM-C can isolate correct anomalous nodes 25 and 24 out of 30 cases respectively. On the other hand, with a simple rule of determining the anomalous node in dataset3 (i.e., finding out the anomalous node that can explain all the failed patterns),  $S^3$  can discover correct anomalous nodes in all 30 cases. Future work will involve a detail comparison study by formalizing the anomalous node identification process from pattern-based evidences especially for multiple anomalous nodes.*

## 5 Conclusions

Based on spatiotemporal causal graphical modelling, this work presents two approaches—the sequential state switching ( $S^3$ ) and artificial anomaly association ( $A^3$ )—for root-cause analysis in complex cyber-physical systems. With synthetic data, the proposed approaches are validated and showed high accuracy in finding failed patterns and diagnose for the anomalous node. Advantages of the proposed methods include -

1. *Ability to handle multiple nominal modes:* The STPN+RBM framework is capable of learning multiple modes as nominal, which corresponds to diverse operation modes in most physical systems (e.g. complex cyber-physical systems). The root-cause analysis approaches proposed in this work enjoy this benefit and are validated with dataset1.
2. *Accuracy:* The proposed approaches— $S^3$  and  $A^3$ —demonstrate high accuracy in root-causes analysis with synthetic data. Moreover, the approach shows considerable credibility in finding out all of the failed patterns while avoiding false negatives.



3. *Scalability*: The approaches are scalable with the size of the system. This is important in complex cyber-physical systems where hundreds of sub-systems are often involved.
4. *Robustness*: Compared with VAR model, the proposed approach can more effectively isolate the fault node with less incorrectly discovered patterns.  $S^3$  method can also find out the maximal set of failed patterns while avoiding false positives.

Future work will pursue: (i) inference approach in node failure including single node and multiple nodes, (ii) detection and root-cause analysis of simultaneous multiple faults in distributed complex systems.

## References

- [1] Tao Yuan and S Joe Qin. Root cause diagnosis of plant-wide oscillations using granger causality. *Journal of Process Control*, 24(2):450–459, 2014.
- [2] Junichi Mori, Vladimir Mahalec, and Jie Yu. Identification of probabilistic graphical network model for root-cause diagnosis in industrial processes. *Computers & Chemical Engineering*, 71:171–209, 2014.
- [3] Rinat Landman, Jukka Kortela, Q Sun, and S-L Jämsä-Jounela. Fault propagation analysis of oscillations in control loops using data-driven causality and plant connectivity. *Computers & Chemical Engineering*, 71:446–456, 2014.
- [4] Ishanu Chattopadhyay. Causality networks. *arXiv preprint arXiv:1406.6651*, 2014.
- [5] Huida Qiu, Yan Liu, Niranjana A Subrahmanya, and Weichang Li. Granger causality for time-series anomaly detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1074–1079. IEEE, 2012.
- [6] Gang Li, Tao Yuan, S Joe Qin, and Tianyou Chai. Dynamic time warping based causality analysis for root-cause diagnosis of nonstationary fault processes. *IFAC-PapersOnLine*, 48(8):1288–1293, 2015.
- [7] Qin Zhang, Chunling Dong, Yan Cui, and Zhihui Yang. Dynamic uncertain causality graph for knowledge representation and probabilistic reasoning: Statistics base, matrix, and application. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(4):645–663, 2014.
- [8] S. Krishnamurthy, S. Sarkar, and A. Tewari. Scalable anomaly detection and isolation in cyber-physical systems using bayesian networks. In *Proceedings of ASME Dynamical Systems and Control Conference, San Antonio, TX, USA*, 2014.
- [9] Chao Liu, Sambuddha Ghosal, Zhanhong Jiang, and Soumik Sarkar. An unsupervised spatiotemporal graphical modeling approach to anomaly detection in distributed CPS. In *Proceedings of the International Conference of Cyber-Physical Systems, (Vienna, Austria)*, 2016.
- [10] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313.5786:504–507, 2006.
- [11] Soumalya Sarkar, Soumik Sarkar, Nurali Virani, Asok Ray, and Murat Yasar. Sensor fusion for fault detection and classification in distributed physical processes. *Frontiers in Robotics and AI*, 1:16, 2014.
- [12] C. Rao, A. Ray, S. Sarkar, and M. Yasar. Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns. *Signal, Image and Video Processing*, 3(2):101–114, 2009.
- [13] Soumik Sarkar, Abhishek Srivastav, and Madhusudana Shashanka. Maximally bijective discretization for data-driven modeling of complex systems. In *American Control Conference (ACC), 2013*, pages 2674–2679. IEEE, 2013.
- [14] Soumik Sarkar and Abhishek Srivastav. A composite discretization scheme for symbolic identification of complex systems. *Signal Processing*, 125:156 – 170, 2016.

- [15] Geoffrey E Hinton. A practical guide to training restricted boltzmann machines. In *Neural Networks: Tricks of the Trade*, pages 599–619. Springer, 2012.
- [16] Kin Gwn Lore, Daniel Stoecklein, Michael Davies, Baskar Ganapathysubramanian, and Soumik Sarkar. Hierarchical feature extraction for efficient design of microfluidic flow patterns. In *Proceedings of The 1st International Workshop on “Feature Extraction: Modern Questions and Challenges”*, NIPS, pages 213–225, 2015.
- [17] Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461. Morgan Kaufmann Publishers Inc., 1996.
- [18] Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 366–374. UAUI Press., 2008.
- [19] Rainer Goebel, Alard Roebroek, Dae-Shik Kim, and Elia Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21(10):1251–1261, 2003.